
Iterative summarization towards natural abstraction understanding¹

Yoann Poupart
ENS Lyon

Abstract

Presentation of the results obtained on iterative summarization. The goal was to train two complementary networks based on GPT-2. The first network is trained to summarize, and the second is trained to exemplify. Both networks are trained and evaluated on the same datasets. With these two networks, it is possible to measure or at least explore the Telephone Theorem. The information diffusion observed in the two networks could give insights into the abstraction capacities and processes of GPT-2.

Keywords: Mechanistic interpretability, ML safety, Summarization, Exemplification, Natural abstraction

Disclaimer: don't expect much of this incomplete work.

1. Introduction

The Natural Abstraction hypotheses can be summarized as follows: abstraction is universal regarding the abstractor and the abstracted [1]. In addition, for this work, the Telephone theorem [2] and the abstraction redundancy [3] are the hypotheses considered.

Even though this work is not directly testing the Natural Abstraction hypotheses, it could be a starting point. For self-improvement and work introspection, I made a few comments related to the different judging criteria.

- ML safety: this work could be informative on the abstraction capacities of neural networks. It could bring insight into deep reasoning processes, but it could also lead to experiments on specification issues.
- Mechanistic interpretability: no real contribution.
- Novelty: nothing regarding the methods or the ideas; just put some pieces together.
- Generality: not really.
- Reproducibility: all the work produced can be found on the associated GitHub repository².

¹ Research conducted at the Apart Research Alignment Jam #4 (Mechanistic Interpretability), 2023 (see <https://itch.io/jam/mechint>)

² GitHub repository link: https://github.com/Xmaster6y/Iterative_summarisation.

2. Methods

The idea was to train two complementary networks based on GPT-2. The first network is trained to summarize, and the second is trained to exemplify. Both networks are trained and evaluated on the same datasets.

- Iterations: the two networks can discuss by iteratively responding to the other network. The goal was here to show the diffusion of information throughout the conversation. The track should be put on the summaries. The hypothesis is that good abstracted ideas by the networks will remain a constant in the different summaries.
- Interpretability tools: the tools used to understand the networks are the ones given in Neel Nanda's library TransformerLens³. Few experiments gave meaningful results; see the results explained below with the associated tools used.

3. Results

The first qualitative results to present are different examples of discussion between the two networks. The generation was constrained to 20 new tokens for the summarizer and 200 new tokens for the exemplifier. Due to these relatively arbitrary (slightly related to the dataset) constraints, the reviews were sometimes too long, and multiple summaries can be produced in a single prediction.

[review]: I love this adaptation of the classic tale. Henry Winkler is, of course, one of my favorite actors. This is a different slant from the original but it gets the message across none the less. Well worth the viewing.

[summary]: Good Adaptation of the Classic Tale

[summary]: Henry WinklerIN

[review]: With some interesting lines this see a fantastic comedy movie! Really comes to life from Tim Hunt. Also but not some movies can be made with behavior.

[summary]: Good movie!

[review]: Excellent movie! A must have classic!

[summary]: # Scrooge!

[review]: More of this one as Is! I loved the original series. Bob Newhart is a nice Scrooge movie show the way "which is the sitcom" rather Hollywood coming to a small screen. Correctly brilliant as it is supposed to allow buddeeling to omit Dr. Slow, Pig and r ___ along with Krule and our grand audience...the bird is not as written at that way and there's not to accomplish what can come out of any of the Great's DVD products out there. The character of Steed may have Aken Doctor who is not and mainly easy to watch - I mostly bebspressed for longer - the most what is the Lazzan Deelee woman who plays Miss Havisham would point out of all people in his best set. And maid Leslie plays Miss Havisham wanna-do, you get to take the key part out of his production. Miss Havisham who isn't more like I did like Newhart

³ GitHub repository link: <https://github.com/neelnanda-io/TransformerLens>

[review]: The counter scene when Allen's character says he is robbing the bank and has a "gub". That is hilarious!! Many more humorous scenes!
 One of America's best comics ever!!
 [summary]: Very Funny
 [summary]: Samantha An old man chick when Ivan allowed to exercise with a huge jump out. A standards.
 [review]: Samantha Fans should love the DVD herself and Jessica Fiennes Wilson.
 [summary]: numerous series. Bob Here's Little Essence
 [review]: this is a superb series and definitely not earlier with your Pelican hair collection. I do not understand why people have at time to this DVD it so much. The characters and acting are incredible. Pardon my audience. I loved them but not to admit how them give their goshog : them to drown. If you are off to the bible so much your barrier to will write much about them to the book. It's more easily written on the whole. There are many BBC set 4 Seasons and so was my ancestors. Pothet (who I winds up at times) to see actors stand up as we had a treat nemesis. Learned how it was difficult to read on the show to keep our head interested in this production way. The English made its appearance much worse on the tape is more offensive than to throw out a dart on into a phonation device that is causing accuracy. I completely turned down the ranking of this one off.
 [summary]: okay.
 [review]: This is a good movie but would the more developed and sophisticated setting are great. Having seen is in a known response to your liking something. I love none of the entre time movies and how they bring us back to the end.

Figure 1 – Two discussions between the trained networks. The first begins with a review of the training dataset and the second with one from the evaluation dataset. In red are the samples from a dataset. The first summary of the discussions (in red) is not in the prompt fed to the model.

The most important interpretations are that first, during reduction from a review to a summary, a lot of information is lost. Then the exemplifier network (and, in a lesser way, the summarizer network) often only repeats existing content. This latter conclusion is corroborated by the copying score results and is mostly due to the size of the summaries fed during fine-tuning. Although some discussions conserve the meaning of the review, it clearly appears that this dataset is not the most appropriate.

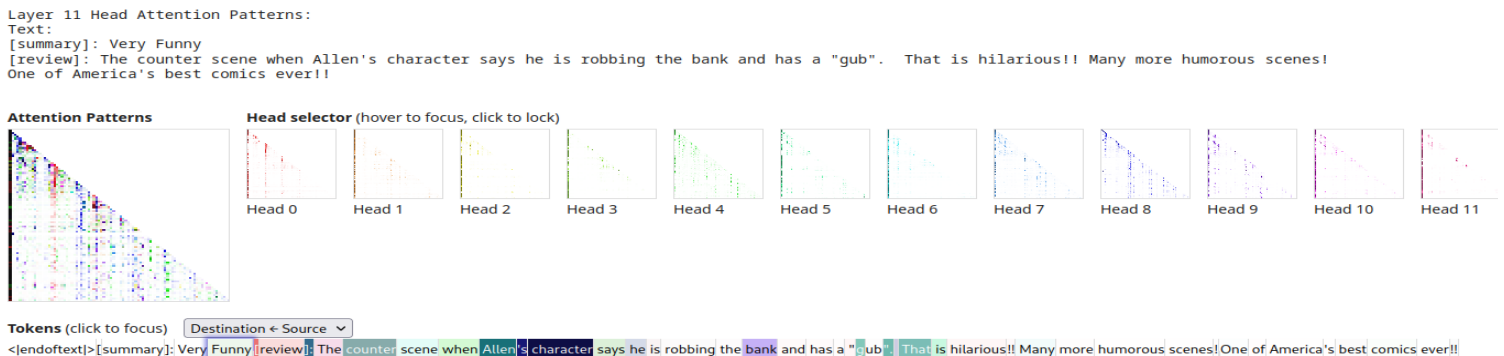


Figure 2 – Patterns related to the word “funny” for the summarizer.

Layer 11 Head Attention Patterns:

Text:

[review]: The counter scene when Allen's character says he is robbing the bank and has a "gub". That is hilarious!! Many more humorous scenes!

One of America's best comics ever!!

[summary]: Very Funny

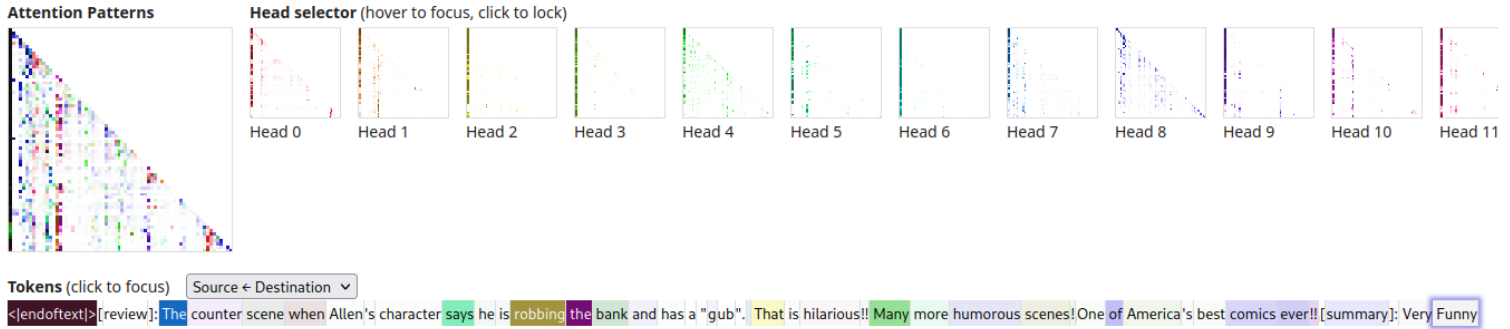


Figure 2 – Patterns related to the word “funny” for the exemplifier.

The results for the attention patterns (figures 2 and 3) show the duplicity of the behaviour of the two networks. While the overall patterns didn’t give much insight when focusing on a particular word it is possible to see the links made by the networks. The asymmetry is also due to order inversion for the review and the summary.

Further experiments could be conducted notably by studying the stream (residual or not) throughout the two networks. It could even be possible to consider a merged network to be trained simultaneously. The big difficulty would be to avoid training a network only by copying its entry. For that, the global loss could be made of the final output loss and an inner loss from the first part of the merged network. It is unsure whether this track will be viable but it could be worth the try.

4. Discussion and Conclusion

The results, although disappointing, could be the consequence of multiple weaknesses. First, obviously with more time and resources, fine-tuning could be done on a bigger model with a bigger dataset. Then, dataset choice and the prompting training are highly debatable to really tackle the notion of abstraction. Finally, the fine-tuning framework was not ideal and could be improved. This work could be extended with natively iterative models proposed in [4] for document summarization, which could be developed to improve abstraction [5].

5. References

- [1] Johnswentworth. (n.d.). *Testing the natural abstraction hypothesis: Project intro*. LessWrong. Retrieved January 22, 2023, from <https://www.lesswrong.com/posts/cy3BhHrGinZCp3LXE/testing-the-natural-abstraction-hypothesis-project-intro>
- [2] Johnswentworth. (n.d.). *The telephone theorem: Information at a distance is mediated by deterministic constraints*. LessWrong. Retrieved January 22, 2023, from

<https://www.lesswrong.com/posts/jJf4FrFiQdDGg7uco/the-telephone-theorem-information-at-a-distance-is-mediated>

- [3] Johnswentworth. (n.d.). *Abstractions as redundant information*. LessWrong. Retrieved January 22, 2023, from <https://www.lesswrong.com/posts/vvEebH5jEvxnJEvBC/abstractions-as-redundant-information>
- [4] Chen, X., Gao, S., Tao, C., Song, Y., Zhao, D., & Yan, R. (2018). Iterative Document Representation Learning Towards Summarization with Polishing. *arXiv*. <https://doi.org/10.48550/arXiv.1809.10324>
- [5] J. Li, C. Zhang, X. Chen, Y. Cao and R. Jia, "Improving Abstractive Summarization with Iterative Representation," *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206950.